A three tier architecture applied to LiDAR processing and monitoring

Efrat Jaeger-Frank[a,*], Christopher J. Crosby[b], Ashraf Memon[a], Viswanath Nandigam[a], Jeffery Conner[b], J. Ramon Arrowsmith[b], Ilkay Altintas[a] and Chaitan Baru[a]

[a]*San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*
Tel.: +1 (858) 822 3694; E-mail: {efrat,amemon,viswanat,altintas,baru}@sdsc.edu

[b]*Department of Geological Sciences, Arizona State University, Tempe, AZ 85287, USA*
Tel.: +1 480 965 3541; E-mail: {chris.crosby, jsconner, ramon.arrowsmith}@asu.edu

Abstract. Emerging Grid technologies enable solving scientific problems that involve large datasets and complex analyses, which in the past were often considered difficult to solve. Coordinating distributed Grid resources and computational processes requires adaptable interfaces and tools that provide modularized and configurable environments for accessing Grid clusters and executing high performance computational tasks. Computationally intensive processes are also subject to a high risk of component failures and thus require close monitoring. In this paper we describe a scientific workflow approach to coordinate various resources via data analysis pipelines. We present a three tier architecture for LiDAR interpolation and analysis, a high performance processing of point intensive datasets, utilizing a portal, a scientific workflow engine and Grid technologies. Our proposed solution is available to the community in a unified framework through a shared *cyberinfrastructure*, the GEON portal, enabling scientists to focus on their scientific work and not be concerned with the implementation of the underlying infrastructure.

Keywords: Scientific worfklows, grids, provenance, LiDAR, grid portals

1. Introduction

With improvements in data acquisition technologies comes an increase in the volume of scientific data. The demand for efficient processing and management of the data have made Grid infrastructures an essential component in a wide range of scientific domains. Grid infrastructure technologies enable large scale resource sharing and data management, collaborative and distributed applications and high performance computing, for solving large scale computational and data intensive problems. However, the distributed and heterogeneous nature of Grid clusters, such as various hardware platforms and software systems, access and interaction interfaces and data and resource management systems, make the Grid environment difficult to use by the layman, and thus require additional management to coordinate the multiple resources. In this paper we propose a coordination of Grid resources, such as data access, analysis and visualization tools in a scientific workflow environment as part of a three tier architecture. The workflow system provides a modularized and configurable environment. It gives the freedom to easily plug-in any process or data resource, to utilize existing sub-workflows within the analysis, and easily extend or modify the analysis using a drag-and-drop functionality through a graphical user interface.

The Geosciences Network (GEON) [19] is an NSF-funded large Information Technology Research (ITR) project to facilitate collaborative, inter-disciplinary science efforts in the earth sciences. GEON is developing an infrastructure that supports advanced semantic-based discovery and integration of data and tools via portals (the GEON portal), to provide unified and authenticated access to a wide range of resources. These

[*] Corresponding author.

ISSN 1058-9244/06/$17.00 © 2006 – IOS Press and the authors. All rights reserved

resources allow geoscientists to conduct comprehensive analyses using emerging web and Grid-base technologies in order to facilitate the next generation of science and education. One of the challenging problems GEON is currently focusing on is distribution, interpolation and analysis of LiDAR (Light Distance And Ranging) [29] point cloud datasets. The high point density of LiDAR datasets pushes the computational limits of typical data distribution and processing systems and makes grid interpolation difficult for most geoscience users who lack computing and software resources necessary to handle these massive data volumes. The geoinformatics approach to LiDAR data processing requires access to distributed heterogeneous resources for data partitioning, analyzing and visualizing all through a single interactive environment. We present a three tier architecture, called the *GEON LiDAR Workflow* (*GLW*) [8] that utilizes the GEON portal as a front end user interface, the Kepler [16] workflow system as a comprehensive environment for coordinating distributed resources using emerging Grid technologies, and the Grid infrastructure, to provide efficient and reliable LiDAR data analysis.

The scientific workflow-based approach already has advantages as workflows enormously improve data analysis, especially when data is obtained from multiple sources and generated by computations on distributed resources and/or various analysis tools. These advances in systematic analysis of scientific information made possible by workflows have unleashed a growing need for automated data-driven applications that also provide scientists with interfaces to design, create, execute, share, reuse scientific workflows, and collect and manage the provenance of the data and processes with little overhead. The scientific workflow approach offers a number of advantages over traditional scripting-based approaches, including ease of configuration, improved reusability and maintenance of workflows and components (actors), automated provenance management, “smart” re-running of different versions of workflow instances, monitoring of long running tasks, and support for fault-tolerance and recovery from failures.

1.1. Related work

Coordinating distributed Grid resources to a scientific workflow has been identified in Several Grid workflow systems. In Kepler [16], Taverna [25] and Triana [26], Grid tools are available wrapped as local components and are coordinated through a user friendly visual programming environment. In Pegasus [4], abstract workflow descriptions are mapped to available Grid resources at execution time. Dealing with massive volumes of data and high performance computations, the need for provenance collection arises. This need has been widely acknowledged and is evident in a numerous applications and systems. The Chimera [14] Virtual Data System uses a process model for tracking provenance in the form of data derivations. In the myGrid project, provenance data is recorded for workflows in XScufl language [18]. The PASOA project [21] aims at providing interoperable means for recording provenance data via an open protocol. In this paper we present a framework for utilizing a back end Grid workflows engine, accessible through a portal front end environment, and a provenance recording system for efficient large scale data processing.

The rest of this paper is organized as follows. Section 2 provides an introduction to LiDAR data and describes the traditional processing approach. Section 3 gives a brief overview of the Kepler scientific workflow system. The novel approach for LiDAR processing, utilizing the Kepler workflow engine through the GEON portal is described and analyzed in Section 4. Section 5 describes the LiDAR monitoring system. We conclude in Section 6 and illustrate the main contributions of this work aimed at making the GEON LiDAR Workflow (GLW) a leading portal for LiDAR processing.

2. Introduction to LiDAR and previous approach

LiDAR (Light Distance And Ranging, a.k.a. ALSM (Airborne Laser Swath Mapping)) data is quickly becoming one of the most exciting new tools in the geosciences for studying the earth’s surface. Airborne LiDAR systems are composed of three separate technologies: a laser scanner, an Inertial Measurement Unit (IMU) and a Global Positioning System (GPS) all configured together to calculate the absolute location for the earth’s surface based upon each individual laser return. The systems typically record one or more returns per square meter on the ground with an absolute vertical accuracy of better than 15 cm. Capable of generating digital elevation models (DEMs) more than an order of magnitude more accurate than those currently available, LiDAR data offers geologists the opportunity to study the processes the shape the earth’s surface at resolutions not previously possible. LiDAR data is currently being utilized by earth scientists for a wide variety of tasks, ranging from evaluating flooding hazards to studying earthquake faults such as the San Andreas [29].

Unfortunately, access to these datasets for the average geoscience user is currently difficult because of the massive volumes of data generated by LiDAR. The distribution, interpolation and analysis of large LiDAR datasets, which frequently exceed a billion data-points, presents significant computational challenges for users who lack the computational resources to handle these types of data volumes. Currently, the popularity and rate of acquisition of LiDAR data far outpaces the resources available for researchers who wish to work with these data.

Typically, geoscientists who wish to work with LiDAR point cloud data are faced with the daunting task of wading through hundreds or thousands of ascii flat files to find the subset of data they are interested in. Once they have selected the piece of point cloud data of interest, they typically must interpolate these data to a regularized grid, called a digital elevation model (DEM), that they can then run their analysis on. The act of interpolating tens of millions of LiDAR data points on a desktop PC with the software packages available and familiar to most earth scientists presents a significant challenge.

The geoinformatics approach to LiDAR processing presented here represents a significant improvement in the way that geoscientists access, interpolate and analyze LiDAR data. The improved, internet-based approach, acts to democratize access to these exciting but computationally challenging data.

3. Kepler: A scientific workflow system

Kepler [1,12] is a cross-project, multi-disciplinary collaboration to build open source tools for scientific workflows that provide domain scientists with an easy-to-use, yet powerful system for capturing and automating their ad-hoc process. Kepler is built on top of the PtolemyII system developed at UC Berkeley, which provides a set of java APIs for modeling heterogeneous, concurrent and hierarchical components by means of various models of computations [3,30]. Kepler provides the scientists with a repetitive and configurable environment available through a graphical user interface and as a command-line tool. It combines high-level workflow design with execution and runtime interaction, access to local and remote data and legacy applications, and local and remote service invocation along with a built-in concurrency control and job scheduling mechanism.

Computational units in Kepler are called *actors*, which are reusable components communicating with each other via input and output ports. The control of flow of actors is orchestrated by a *director* that specifies the model of computation. Kepler uses the Modeling Markup Language (MoML), inherited from the underlying PtolemyII system, as its workflow description language. MoML is a modularized and extensible XML modeling language where actors can be defined as place holder stubs to be set prior to the workflow execution.

In Kepler we have created a highly configurable provenance component, called *Provenance Recorder* (*PR*), that is used to mine data from a workflow run. The PR is consistent with Kepler's actor oriented paradigm and can be easily added to a workflow through the user interface. We have also added a *Smart Rerun Manager* (*SRM*) to enable efficient reruns of a workflow, by mining the stored provenance data. Thus when re-executing a workflow or a parameterized modification of a workflow, "redundant"/previously mined unaffected parts of the workflow need not be re-executed. The data products generated in previous runs are used as inputs to the actors that are to be rerun. Additional information about the Kepler PR and SRM is available at [13]. In the next sections we describe how utilizing the Kepler features enhances LiDAR processing and monitoring.

4. A scientific workflow based approach for LiDAR processing

In the following section we present a three tiered Kepler scientific workflow solution for enhancing distribution, interpolation and analysis of LiDAR datasets.

4.1. Coordinating distributed resources in a single environment

LiDAR processing requires three main computational steps each deployed on a distributed resource: querying point cloud datasets, processing the data using various interpolation algorithms, and visualizing the results. Coordinating these steps in a Kepler scientific workflow provides scheduling and monitoring of each task and communication between the various resources. Furthermore, the scientific workflow environment gives us modularity and extensibility through

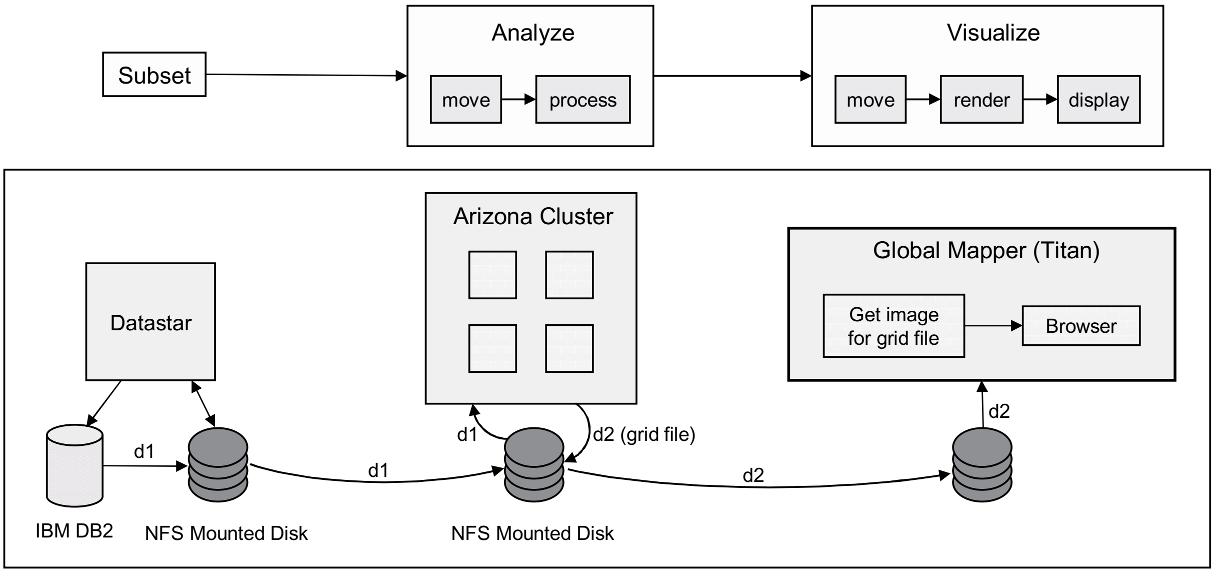
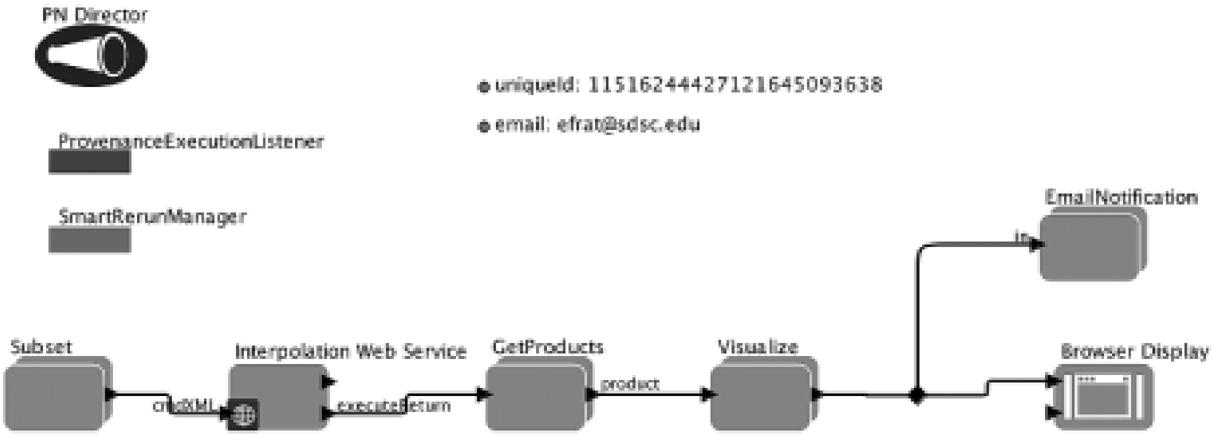


Fig. 1. A *subset*, *analyze*, *visualize* workflow pattern, coordinating distributed heterogeneous resources in a single scientific workflow environment. Each computation is performed under a diverse authority and intermediate data products (d_1, d_2) are passed between compute nodes.



● uniqueId: 11516244427121645093638
● email: efrat@sdsc.edu

Fig. 2. An executable workflow instance example created on the fly from the workflow template based on the user selections.

reusable building block components. The LiDAR processing workflow, depicted in Fig. 1, provides a *conceptual* workflow where each of the components can be dynamically customized by the availability of the data and processing algorithms and is set on the fly prior to the workflow execution (Fig. 2 gives a snapshot of the top level of an instantiated workflow instance). This modularized approach can be captured as a workflow pattern of *subset*, *analyze*, and *visualize*. Below we elaborate on each of the processing steps.

The massive amount of LiDAR data (currently two datasets at ∼1 billion points each) were uploaded and organized in a DB2 spatial database on *DataStar* to provide a unified structure to the collected data. The subset query returns all points (X,Y,Z) that reside within a user selected bounding box. The database connection information along with the query are specified as workflow parameters and are set on the fly prior to the workflow execution. The query is then performed on DataStar where the result is also stored on an NFS mounted disk.

The analysis step consists of an interpolation algorithm. As shown in Fig. 1, the query response is shipped to the *analysis* cluster and is then interpolated into a regularized grid. Currently we use the GRASS spline interpolation [11] algorithm deployed on the Arizona GEON four nodes cluster. Other interpolation algorithms, for example, Inverse Distance Weighted (IDW) or Kriging algorithms may be plugged in as well. Interpolation of the high-point density LiDAR data currently constitutes the bottleneck in the overall process. Parallelization of the interpolation code or alternative interpolation algorithms will likely result in improved performance.

The interpolation results may be visualized and/or downloaded. At present, the Global Mapper imaging tool [10] is used to create a display from the interpolated results. Global Mapper, available to the GEON community through a web service, takes an ASCII grid file as an input and returns a URL to the resulting image. The image can be displayed on a web browser and requires no specific visualization components. Other visualization methods may be applied as well.

4.2. A three tier architecture

The GEON project is developing a web-based portal for gathering geoscience applications and resources under a single roof. In order to make the workflow based approach for LiDAR processing uniformly accessible through the GEON portal, our proposed solution is based on a three tier architecture: the *portal layer*, the *workflow layer* or control layer, which is the Kepler workflow system, and the *Grid layer* as the computation layer. The *portal layer*, a portlet, serves as a front end user interface available from the GEON portal. The portal's underlying GAMA [6] authentication infrastructure provides a role based authentication to resources available through the GEON portal using a single sign-on. The GLW portlet enables the user to partition the LiDAR data using an interactive mapping tool and attribute selection through a WMS map [20]. Algorithms, processing attributes and desired derivative products, are also chosen using a web interface. Within Kepler, one can design a predefined parameterized workflow template which is modularized and configurable using place holder stubs to be set prior to the workflow execution. The aforementioned *"subset, analyze, visualize"* workflow pattern serves as a conceptual workflow template, defined in the Kepler workflow description language, MoML. A workflow instance is created on the fly from the conceptual workflow based on the user selections. The instantiated workflow is then scheduled to be executed by the workflow layer.

The *workflow layer*, also referred to as the main control layer, communicates both with the portal and the Grid layers. This layer, controlled by the Kepler workflow manager, coordinates the multiple distributed Grid components in a single environment as a data analysis pipeline. It submits and monitors jobs onto the Grid, and handles third party transfer of derived intermediate products among consecutive compute clusters, as defined by the workflow description. In addition, it sends control information (a.k.a. tokens) to the portal client about the overall execution of the process. The workflow is executed by the Kepler engine in a batch mode. Once a job is submitted, the user can detach from the system and receive an email notification after the process has completed.

As the LiDAR processing workflow involves long running processes on distributed computational resources under diverse controlling authorities, it is exposed to a high risk of component failures, and requires close monitoring. In order to overcome these failures with minimal user involvement, Kepler provides a data provenance and failure recovery capability by using a job database and smart reruns. The job database is used for logging the workflow execution trace and storing intermediate results along with the associated processes/components that were used to produce them. The workflow engine maintains information about the status of each intermediate step, and this can be used to initiate a smart rerun from a failure point or a checkpoint. These advanced features thus eliminate the need to re-execute computationally intensive processes. The LiDAR monitoring system is further described in Section 5.

The *Grid layer*, or the execution layer is where the actual processing implementations are deployed on the distributed computational Grids. Currently a simple submission and queueing algorithm is used for mapping jobs between various resources based on the number of allocated tasks and the size of the data to be processed. In the near future we plan to utilize the Pegasus [15] Grid scheduler to benefit from mapping jobs based on resource efficiency and load, thus making the process more robust. We also plan to extend this further by deployment of the computationally challenging processes on a higher performance machine, for example, DataStar. DataStar [24] consists of 32 P690 processors with 128 GB of memory running at 1.7 GHz, each with a gigabit connection to an underlying SAN disk infrastructure, making it an ideal machine for compute intensive tasks.

Figure 3 depicts the overall architecture design.

4.3. Data upload and access

As initial pilot datasets, LiDAR data collected along the Northern San Andreas Fault (NSAF) in Sonoma

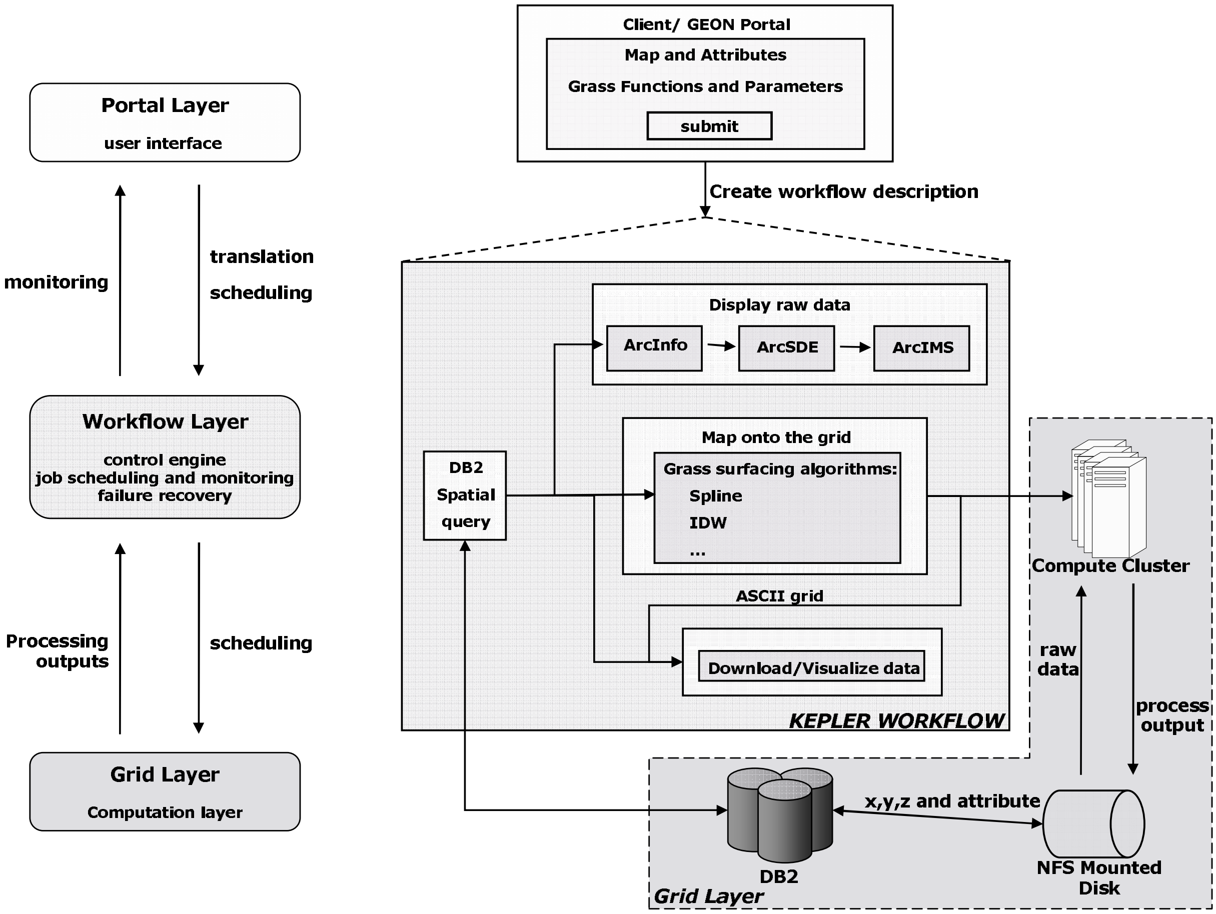


Fig. 3. Three tier LiDAR processing architecture depicting the portal, workflow and Grid layers.

and Mendocino Counties, California and data from the Western Rainier Seismic Zone in Pierce County, Washington were utilized. The approximately 1.2 billion points (which is about 6 GB of data) in each of these datasets were stored in a DB2 Spatial Extender Relational Database on DataStar. To optimize query performance, and due to its sheer volume, the data were grouped by USGS Quarter Quadrants [27] and organized in the database as a table per quarter quadrant. The multiple data segments are accessible through the WMS map using a metadata table that specifies the bounding box coordinates for each segment. Each record in the datasets consists of a coordinate, elevation and corresponding attributes. The Spatial Extender feature in DB2, leveraging the power of standard SQL for spatial data analysis, enables storage, access and management of the spatial data. It allows users to choose from a set of standard projection planes as well as define custom projection parameters. Furthermore, in addition to standard database indexing, it provides a grid indexing technology for indexing multi-dimensional spatial data, dividing a region into logical square grids, thus facilitating region based subset queries. Taking advantage of these features, the LiDAR NSAF coordinate data was converted into spatial geometry points using the appropriate projection plane (CA State Plane zone II, NAD83 datum).

4.4. Analysis of the approach

The three tier based architecture utilizes the Kepler workflow system to reengineer LiDAR data processing from a geoinformatics approach. The data and processing tools are accessible through a shared infrastructure and are coordinated using the Kepler engine. Kepler provides an incremental design and development environment which allows users to create incomplete workflow templates, which can be filled on demand prior to the workflow execution. Seamless access to resources and services is enabled through existing generic components such as SSH, Web Service, database access and command line processor actors. These generic building block components offer the flexibility to plug-in any applicable data or process, thus providing a customizable and modularized environment. The GLW currently has access to two datasets both stored in a DB2 database on DataStar. Other databases sharing a similar structure may be used simply by pointing to a different database location. Currently we use GRASS spline interpolation, available via a web service deployed on the Arizona GEON cluster. Other interpolation algorithms such as GRASS IDW can be applied by updating a WSDL URL parameter. The visualization tools are interchangeable as well. Global Mapper can be replaced with FlederMaus [7] or ArcIMS [5] visualization tools, both deployed on the GEON portal as web services.

The GLW's main processing bottleneck is in the interpolation of the high point density datasets. Currently, the GRASS spline interpolation is limited to processing 1,600,000 points. This problem is being addressed by the GRASS development community and we anticipate a solution to this problem in the near future. Ultimately however, to improve performance, a parallel interpolation algorithm is required along with deployment on a higher performance machine. We are also testing the Pegasus system [4,15] in coordination with the Pegasus group at the University of Southern California for a more efficient mapping of the interpolation sub-workflow onto the Grid.

The workflow solution, though initially designed for LiDAR processing is also applicable in other domains. In the GEON project it is used in gravity modelling, comparing between synthetic and observed gravity models. Both workflows share similar "traits" of *subset*, *interpolate* and *visualize* and thus the same approach can be used.

5. LiDAR monitoring system

5.1. System overview

As previously mentioned above, the LiDAR processing workflow depends on distributed computational resources, and is thus at a high risk of component failures. These failures vary from network failures and database performance issues to system backup and maintenance down time. Another risk is posed by the high performance, long running processes that are also subject to failures. In order to reduce the impact of these failures, the GLW is provided with a LiDAR monitoring system. The monitoring system uses the Kepler provenance collector component, called Provenance Recorder (PR). The PR provides the functionality to create and maintain associations between workflow inputs, workflow outputs, workflow definitions, and intermediate data products. This collected information is used for tracking the history of submitted jobs. The intermediate results along with the other provenance data are then used for debugging purposes and to perform a "smart" rerun of a job. In addition, the monitoring system also maintains the relationship between subsequent jobs and can be used to track job evolution.

In the GLW, provenance tracking stores information to help the user understand how the job ran and what parameters and inputs were associated with the job. The LiDAR monitoring system uses this information to provide a unified framework through the GEON portal for users under their own user space called myGEON to track the history and status of their submitted jobs. A user submits a job, the system records all the job configuration parameters and inputs into the LiDAR jobs database, which is a relational database. Using the Kepler PR, intermediate and end products are recorded and stored in the *SRB* [2]. The user can monitor the status of the job while it is being executed, reproduce the results, and also fine tune the job and resubmit. In case of a failure, the user is able to get feedback on the cause of the error and resubmit the (possibly modified) job. The system also uses the Kepler Smart Rerun Manager (SRM) to initiate a smart rerun from a failure point or a checkpoint and avoid re-execution of high performance processes. The SRM is also used when retrieving processing results. The system first checks whether the interpolation response is already available on the SRB, in which case it invokes only the visualization part of the workflow using a smart rerun. If the interpolation response is not available, the SRM will be used to invoke a smart rerun from a certain available checkpoint (Fig. 2 illustrates how this functionality is incorporated into a specific workflow).

5.2. Implementation details

The monitoring system relies on two essential Kepler components: The Provenance Recorder (PR) and the Smart Rerun Manager (SRM). Before going into the details of these component, we introduce the *SRB* storage system and the underlying algorithm behind the Smart Rerun Manager named *Vistrail*.

5.2.1. *The Storage Resource Broker* (*SRB*)

SRB [2] is a storage management system designed for the Data Grid environments. SRB provides secure and optimized file transfer functionalities including transparent data replication; archiving, caching, and backup. Using logical spaces, SRB grants a heterogeneous storage mechanism for seamlessly accessing data from various physical resources. Moreover, SRB offers bulk data ingestion, version control and Metadata query functionalities through the mean of MetaData Catalog (MCAT).

In the LiDAR monitoring system the SRB serves as a cache where intermediate workflow products are stored. These products can be retrieved, instead of being recreated, when rerunning a job or submitting a similar job (these similarities stem from same selection area and classification attributes). By using the SRB logical namespace we need not be concerned with how and where the actual physical file is being stored. Furthermore, the SRB metadata functionality allows users to annotate stored data and query for datasets based on these annotations. In the future we plan to exploit this functionality to perform ontological based searches for workflow products, by not only querying based on a logical file name, but also querying based on the user selected bounding box, selected algorithms and other parameters.

5.2.2. *VisTrails*

The VisTrails system [17,23,28] was developed to facilitate system independent, interactive multiple-view visualizations by providing a general infrastructure to create and maintain visualization workflows as well as to optimize their execution. The VisTrails system collects and maintains a detailed provenance record for a workflow run as well as across different versions/instances of a workflow, thus tracking the workflow evolution process. The VisTrails algorithm takes a graph representation of the workflow and searches for sub-graphs that need not be re-executed and thus can be eliminated. The precondition for elimination of these sub-graphs is that the actors they contain have already been run with the current parameters and input data, and their intermediate data products have been stored in the provenance cache.

The LiDAR monitoring system uses the VisTrails algorithm as part of the Smart Rerun Manager (SRM). The SRM analyzes this graph to detect sub-graphs that have been successfully computed before and the sub-graphs that must be rerun. The successfully precomputed sub-graphs are replaced by actors that stream the available intermediate data products from the provenance cache, thus reducing execution time and expended resources.

5.2.3. *The Kepler Provenance Recorder* (*PR*)

The Provenance Recorder (PR) is part of the Kepler provenance framework that enables provenance collection in a workflow instance. The PR collects information that is generated during a workflow run. It is notified whenever a new data product is created, and it then associates the appropriate data lineage information with the data product and puts it in the provenance store. Using the PR gives us information about the specific workflow processing steps, such as which actors were executing on which inputs to generate a certain output.

The location where the PR stores the provenance data is specified by a parameter. This location can vary between a relational database, the SRB and even a flat ASCII file. The PR also provides the functionality to control the amount of collected provenance data during a workflow run by defining the level of detail in the output. Furthermore, the PR can also be selective in which component's intermediate products will be collected. This is very useful as some workflows create a massive number of intermediate data products, which are not always necessary to recreate the results of a certain workflow.

In the GLW, the PR is used to store intermediate and end products of a job run in the SRB. This data is applicable when viewing a workflow's result or issuing a smart rerun based on a previous run, and is also useful for debugging purposes.

5.2.4. *Smart Rerun Manager* (*SRM*)

Kepler provides the functionality to enable efficient reruns of a workflow by mining stored provenance data. The idea behind a smart rerun is as follows. When a user changes a parameter of an actor and runs the workflow again, re-executing all the preceding unchanged steps (actors) in the workflow may be redundant and time consuming. A smart rerun of the workflow takes data dependencies into account and only execute those parts of the workflow affected by the parameter change. The SRM first converts the workflow description, MoML, into a graph, then it uses the VisTrails algorithm to detect previously run sub-graphs that could be eliminated. The ability to store and mine provenance data is necessary in order to enable smart reruns since the intermediate data products generated in previous runs are used as the inputs to the actors that are about to be rerun.

The main bottleneck in the LiDAR processing workflow is in the time consuming, high performance interpolation algorithm. Often a user may just be interested in retrieving different output formats for the same selection area and thus there is no need to re-execute the interpolation process. The mined results are used to create the various formats. The LiDAR monitoring system uses the VisTrails algorithm through the SRM to track whether the same query, with the same parameters, was already issued and if the interpolation response is available by querying the SRB using logical file names. If those are available the SRM reruns only the visualization part of the workflow.

5.3. Advantages

The LiDAR monitoring system provides an interface, that is useful to both the user and the workflow developer, to track the history of all submitted jobs. In the following section we list additional benefits of the system.

5.3.1. Job management and sharing

Using the monitoring system, users are able to manage their jobs through their own user space. This provides a unified interface for users to follow up on the status of their submitted jobs. The system allows them to view the metadata of the job, zoom to a specific bounding box location, track errors, modify a job and resubmit, and view the processing results. The user can register desired workflow products within the GEON portal and easily share it with colleagues. Registering results may also be useful when publishing by being able to refer the reader to interesting results.

5.3.2. Job evolution provenance

Scientific problem solving is an evolving process. In the GLW, a scientist may start with a specific area of interest, update parameters and attributes, fine-tune the selected bounding box, re-iterate the process with additional algorithms/output formats, and compare between various selection areas until a satisfying result has been reached. This repeating process provides information on the steps leading to the solution and end results. Maintaining the relationship between subsequent jobs provides the Job Evolution Process. The monitoring system, through provenance tracking, provides the functionality to link between related jobs, that is, jobs that were submitted as a fine-tuning/update of a previously run job. The job evolution provenance can be viewed as a tree of related jobs. Using it the user can follow up on and compare between related jobs and products. It is especially useful in publications. Often, the scientist may want to track the specific steps that led to a specific configuration for validation and reproducibility purposes, and to be able to go back to a certain update point and fine-tune from there.

5.3.3. Future performance improvement

The LiDAR monitoring system uses the SRM to invoke a smart rerun using previously computed products. We also plan to create a repository of job results, using provenance data from previous executions, by storing the interpolation products per bounding box and attributes selection. These would be replaced whenever a containing bounding box, with the same attributes, is selected. Each product will be annotated with metadata information and stored in the SRB, and matching products would be searched for using the SRB metadata query functionality. We would then like to use ArcIMS [5] tools to access contained bounding box with the ones that were already pre-computed. Thus eliminating the need to re-execute the high performance, long running process and improve the workflow's performance.

5.3.4. Usage statistics

The monitoring system appears to be very useful in the GLW development as well. Since the system records all the “traffic” that goes through the GLW, it provides the development team with information about the popularity of the workflow and statistics about its usage. These include the number of page hits, job submissions, and preferred datasets and areas of interest, enabling us to improve the workflow based on the scientists preferences. It can also be used for gathering information about the overall system overhead, timing, load, etc. Furthermore, it provides a venue to track errors and provide background transparent maintenance.

6. Conclusion

In this paper we describe a three tier architecture for LiDAR data processing and monitoring using a comprehensive workflow system, a shared cyberinfrastructure and evolving Grid technologies. We present a novel approach of utilizing a scientific workflow engine as a server side middleware for communicating between a front end web interface and Grid resources.

Furthermore, the GLW uses a multipurpose workflow template which is instantiated on the fly based on the user selections. The new addition to the GLW is the LiDAR monitoring system. The system is an essential component to help users track, in an organized unified framework, the history of their submitted jobs and issue a smart rerun based on previously run jobs. It is especially useful to the heavy users who submit jobs on a daily basis.

This work presents a significant improvement in LiDAR data interpolation and analysis. The first version of this effort is available at the GEON portal [9], and has been incorporated as a public tool. The architecture was already adopted by other projects (for example ROADNet [22]). We plan to further improve the overall performance of the GLW with advanced processing tools such as parallel interpolation algorithms and enhanced visualization methods. Our goal is to provide a centralized location for LiDAR data access and interpolation that will be useful to a wide range of earth science users.

Acknowledgements

GEON is supported by NSF ITR 0225673. Special thanks to Dogan Seber and Ghulam Memon for their support in this work.

References

[1] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jäger-Frank, M. Jones, E.A. Lee, J. Tao and Y. Zhao, *Scientific Workflow Management and the Kepler System*, Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows, to appear, 2005.

[2] C. Baru, R. Moore, A. Rajasekar and M. Wan, The SDSC Storage Resource Broker, Proc. CASCON'98 Conference , Nov.30-Dec.3, 1998, Toronto, Canada.

[3] E.A. Lee et al., PtolemyII Project and System. Department of EECS, UC Berkeley, http://ptolemy.eecs.berkeley.edu/ptolemyII.

[4] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, S. Patil, M. Su, K. Vahi and M. Livny, *Pegasus: Mapping Scientific Workflows onto the Grid*, Across Grids Conference 2004, Nicosia, Cyprus.

[5] ESRI ArcIMS: www.esri.com/arcims.

[6] F. Sacerdoti, S. Chandra and K. Bhatia, *Grid Systems Deployment and Management Using Rocks*, in IEEE Clusters, 2004.

[7] FlederMaus: An interactive 3D visualization system: http://www.ivs3d.com/products/fledermaus/.

[8] The GEON LiDAR data processing: http://www.geongrid.org/science/lidar.html.

[9] The GEON portal: https://portal.geongrid.org:8443/gridsphere/gridsphere.

[10] Global Mapper: http://www.globalmapper.com/.

[11] H. Mitasova, L. Mitas and R.S. Harmon, 2005, Simultaneous Spline Interpolation and Topographic Analysis for LiDAR Elevation Data: Methods for Open Source GIS, *IEEE GRSL* **2**(4), 375–379.

[12] I. Altintas, C. Berkley, E. Jäger, M. Jones, B. Ludäscher and S. Mock, *Kepler: Towards a Grid-Enabled System for Scientific Workflows*, in the Workflow in Grid Systems Workshop in The Tenth Global Grid Forum, Germany, 2004

[13] I. Altintas, O. Barney and E. Jaeger-Frank, *Provenance Collection Support in the Kepler Scientific Workflow System*, IPAW2006, Chicago, Illinois, May 2006.

[14] I. Foster, J. Voeckler, M. Wilde and Y. Zhao, *Chimera: A Virtual Data System for Representing, Query-ing, and Automating Data Derivation*, In Proceedings of the 14th Conference on Scientific and Statisti-cal Database Management, 2002.

[15] J. Blythe, S. Jain, E. Deelman, Y. Gil, K. Vahi, A. Mandal and K. Kennedy, *Task Scheduling Strategies for Workflow-based Applications in Grids*, CCGrid 2005.

[16] Kepler: An Extensible System for Scientific Workflows, http://kepler.ecoinformatics.org.

[17] L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva and H. Vo, *Vistrails: Enabling interactive multipleview visualizations*, In IEEE Visualization 2005, 2005, 135–142.

[18] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau and T. Oinn, *Provenance of e-Science Experiments – experience from Bioinformatics*, In Proceedings of The UK OST e-Science second All Hands Meeting 2003 (AHM'03).

[19] NSF/ITR: GEON: A Research Project to Creat Cyberinfrastructure for the Geosciences, www.geongrid.org.

[20] OpenGIS Web Mapping Specification: http://www.opengeospatial.org/standards/wms.

[21] P. Groth, M. Luck and L. Moreau, *A protocol for Recording Provenance in Service-Oriented Grids*, In Proceedings of the 8th International Conference on Principles of Distributed Systems (OPODIS'04), 2004.

[22] ROADNet: Real-time Observatories, Application and Data management Network: http://roadnet.ucsd.edu.

[23] S. Callahan, J. Freire, E. Santos, C. Scheidegger, C. Silva and H. Vo, *Managing the Evolution of Dataflows with VisTrails*, In Proceedings of the IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow 2006).

[24] SDSC User Services: DataStar: http://www.sdsc.edu/user_services/datastar/.

[25] The Taverna Project: http://taverna.sourceforge.net.

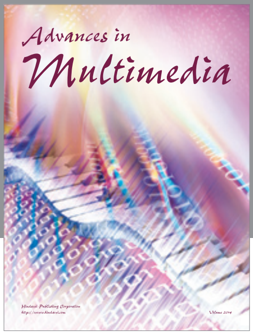
[26] The Triana Project: http://trianacode.org.

[27] USGS Quarter Quadrants: http://www.kitsapgov.com/gis/metadata/support/qqcode.htm.

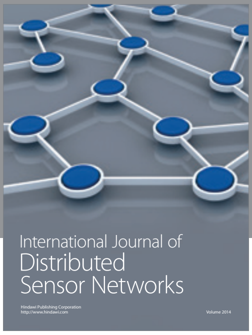
[28] The Visualization Toolkit (VTK), See Website: http://public.kitware.com/VTK/.

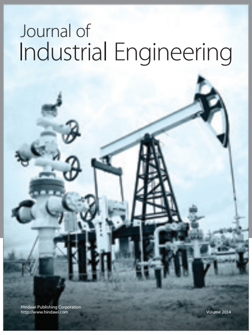
[29] W.E. Carter, R.L. Shrestha, G. Tuell, D. Bloomquist and M. Sartori, 2001, *Airborne Laser Swath Mapping Shines New Light on Earth's Topography: Eos* (*Transactions, American Geophysical Union*) **82** 549.

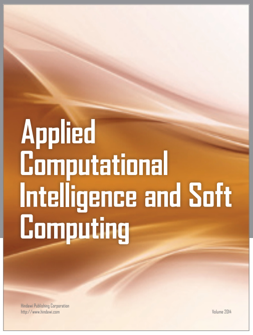
[30] X. Liu, J. Liu, J. Eker and E.A. Lee, *Heterogeneous Modeling and Design of Control Systems*, in Software-Enabled Control: Information Technology for Dynamical System, Tariq Samad and Gary Balas, Wiley-IEEE Press, 2003.

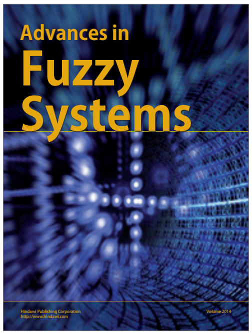

Advances in
Multimedia

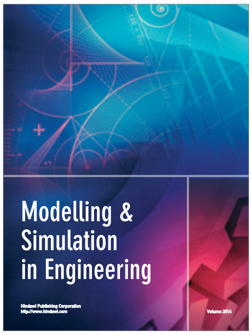
The Scientific
World Journal

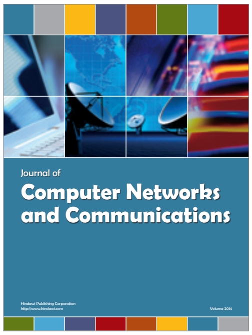

International Journal of
Distributed
Sensor Networks

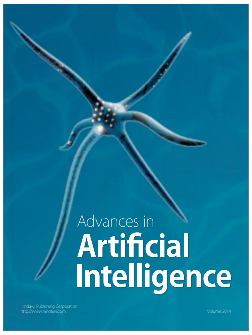

Journal of
Industrial Engineering

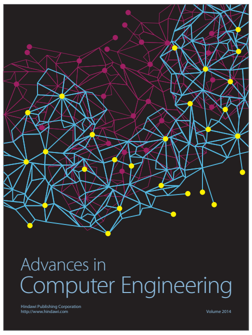

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

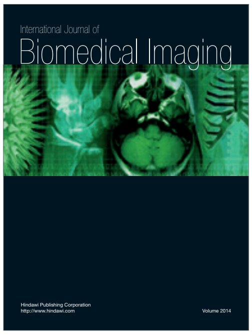

Modelling &
Simulation
in Engineering

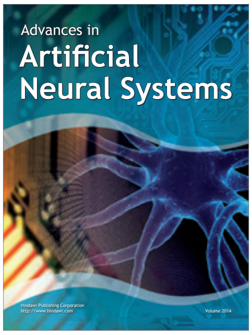

Journal of
Computer Networks
and Communications

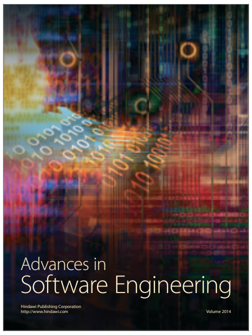

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

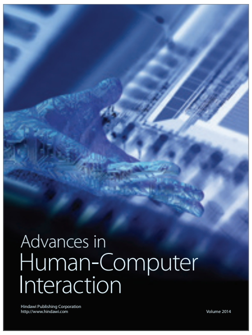
International Journal of
Computer Games
Technology

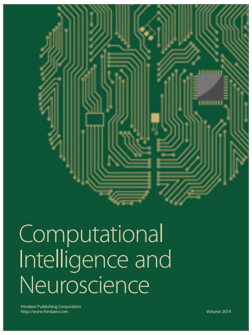

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering